Does It Look Right?

Cultural marking and representational fidelity in text-to-image generators

Jordan Legesse | Supervised by Prof. Amar Ahmad | New York University Abu Dhabi

LAIDLAW FOUNDATION

جامعة نيويورك ابوظبي
NYU ABU DHABI

The question

Ask a model for “a community celebration” in Lyon, Osaka or Lagos. What does it add, and what does it leave out?

Earlier audits compare whole continents. We split them into **21 cultural sub-regions** and ask: does pooling hide differences, does a cheap prompt fix help, and do bias metrics agree?

Study scale

15,120 images generated

21 cultural sub-regions

6 macro-regions

2 open-weight generators

How we measured it

Prompts 5 everyday scenes + 9 cultural categories

↓

4 conditions Neutral · region named · contrastive · CLIP re-rank

↓

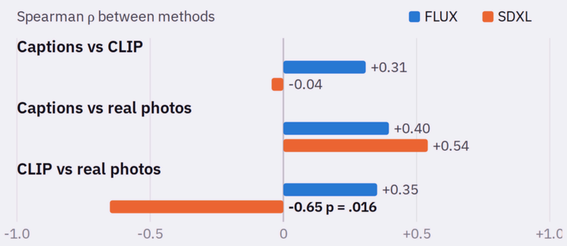
Generators FLUX.1-schnell and SDXL, 20 images per cell

↓

Five lenses

Caption markers: Qwen3-VL captions × 80-term lexicon
Vocabulary: TF-IDF / LDA per cell
Embedding: CLIP similarity to Neutral
Real photos: UCOGC own-country match
Agreement: Spearman ρ across lenses

Do the methods agree?



Only 1 of 6 pairs is significant. Caption, embedding and photo metrics measure related but different things.

Same prompt, different worlds

Neutral prompt no place named
white, woman, wearing, shirt, dark

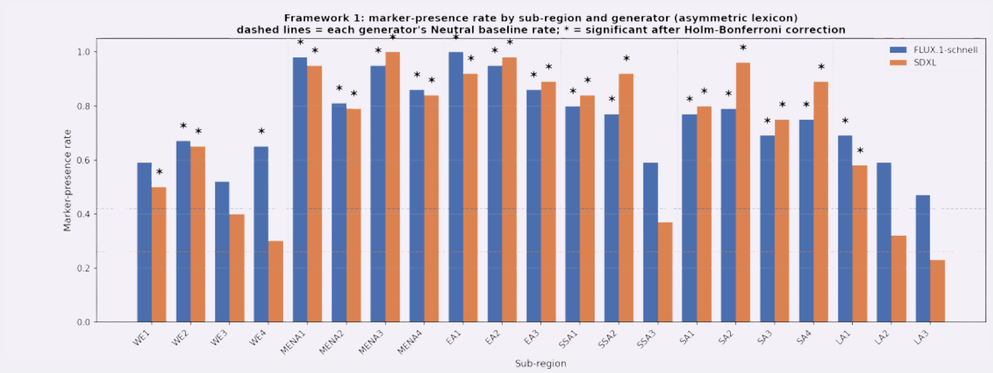
Latin Europe anchor: France
white, red, woman, man, blue

Japanese sphere anchor: Japan
traditional, red, japanese, lanterns, white

Indo-Gangetic anchor: India
woman, red, gold, sari, right

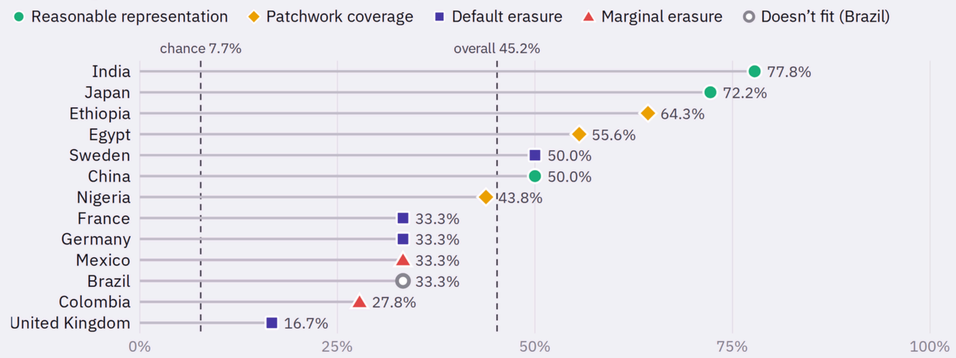
Top caption words for “a community celebration” (FLUX.1-schnell). Western prompts read like the Neutral one; others gain markers.

Non-Western sub-regions are heavily marked; Europe reads as default



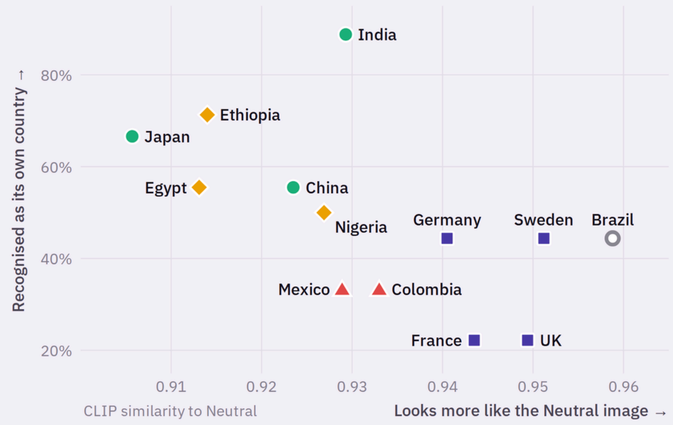
32 of 42 comparisons differ from Neutral after Holm–Bonferroni; none of 84 is statistically equivalent (TOST). Brazil (Latin America 3) is the exception. Markers concentrate in celebration scenes; dinner, commute and play scenes carry almost no regional vocabulary.

Recognisable as itself? Only 45% of the time



Share of generated images closest to their own country’s real photographs (13 countries, both generators pooled). **64.5%** of these decisions are close calls (margin < 0.03). Architecture is the most distinctive category (75%); craft is at chance (7.7%).

Two kinds of erasure



SDXL, 13 sub-regions with real-photo references. ρ = −0.652, p = .016. Colours show the paper’s qualitative taxonomy.

■ Default erasure
Western Europe is the unmarked frame. Its four sub-regions are indistinguishable from one another.

▲ Marginal erasure
Mexico and Colombia get thin, fragile coverage: 83% of their decisions are close calls.

◆ Patchwork coverage
MENA and Sub-Saharan Africa are heavily marked, with a narrow shared vocabulary.

● Reasonable
East and South Asia: distinct vocabulary per sub-region; India matches 78%.

A prompt fix disrupts, but doesn’t redirect

4.3–7.3×
odds of cultural markers with contrastive prompting (p < .001)

p = .34
no reliable move toward the region’s own real target

Change in CLIP similarity to each sub-region’s own target (both generators pooled). Red: moved away. Green: moved closer.

Limitations

- Captions and CLIP ViT-B/32 may carry their own bias (captioner κ = 0.52–0.56)
- Two open-weight models; real photos for 13 of 21 sub-regions
- The taxonomy is a qualitative synthesis; Brazil doesn’t fit it